

# An inequality for the Fourier spectrum of parity decision trees

Eric Blais\*  
University of Waterloo  
eric.blais@uwaterloo.ca

Li-Yang Tan†  
Simons Institute, UC Berkeley  
liyang@cs.columbia.edu

Andrew Wan‡  
Institute for Defense Analyses  
atw12@columbia.edu

June 4, 2015

## Abstract

We give a new bound on the sum of the linear Fourier coefficients of a Boolean function in terms of its parity decision tree complexity. This result generalizes an inequality of O’Donnell and Servedio for regular decision trees [OS08]. We use this bound to obtain the first non-trivial lower bound on the parity decision tree complexity of the recursive majority function.

## 1 Introduction

In this note, we explore connections between two different notions of complexity of Boolean functions  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ : its decision tree complexity, and the sum of its linear Fourier coefficients.

Decision trees are full binary trees with internal nodes labelled by variables  $x_i$  for some  $i \in [n]$  and with leaves labelled with constants  $\ell \in \{-1, 1\}$ . A decision tree  $D$  is said to compute  $f$  if the path from the root to a leaf in  $D$  defined by  $x$  leads to a leaf labelled by  $f(x)$  for every  $x \in \{-1, 1\}^n$ . The *depth* of a decision tree is the maximum number of internal nodes along any root-to-leaf path, and the *decision tree (depth) complexity* of a function  $f$  is the minimum depth of any decision tree  $D$  that computes  $f$ .

Every Boolean function  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  has a unique representation as a multilinear polynomial

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x)$$

where  $\chi_S(x) := \prod_{i \in S} x_i$  and the numbers  $\hat{f}(S) = \mathbf{E}[f(\mathbf{x}) \chi_S(\mathbf{x})] \in [-1, 1]$  are the *Fourier coefficients* of  $f$ . The Fourier coefficients corresponding to singleton sets  $S = \{i\}$ ,  $i \in [n]$  are called *linear Fourier coefficients*. For notational clarity, we will write  $\hat{f}(i)$  to denote the linear Fourier coefficient  $\hat{f}(\{i\})$ . As mentioned above, we consider the measure of complexity of  $f$  determined by the sum  $\sum_{i=1}^n \hat{f}(i)$  of its linear Fourier coefficients.

---

\*Part of this research was done while supported by a Simons Postdoctoral Fellowship at MIT.

†Supported by NSF grants CCF-1115703 and CCF-1319788. Part of this research was done while visiting Carnegie Mellon University.

‡Part of this research was done while visiting Harvard University and supported by NSF grant CCF-964401.

In [OS08] O’Donnell and Servedio established a connection between these two measures of complexity by establishing the following inequality on the linear Fourier coefficients of a Boolean function computed by a depth- $d$  decision tree:

**O’Donnell–Servedio Inequality.** *Let  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  be computable by a decision tree of depth  $d$ . Then  $\sum_{i=1}^n \widehat{f}(i) \leq \sqrt{d}$ .*

In addition to being a natural statement relating a combinatorial notion of complexity (decision tree complexity) to an analytic one (the sum of linear Fourier coefficients), this inequality is also the crux of the main algorithmic result of [OS08], the first algorithm for PAC learning the class of monotone functions to high accuracy from uniformly random labelled examples, running in time polynomial in a reasonable complexity measure of the target function (in this case, its decision tree complexity). To date this remains our best progress towards the goal of efficiently learning monotone polynomial-sized DNFs, a longstanding open problem in PAC learning [Blu03].

## 1.1 Our main result

Another notion of complexity of Boolean functions related to decision trees is their parity decision tree complexity. *Parity decision trees* (PDTs) are generalizations of decision trees where internal nodes are now labelled by subsets  $S \subseteq [n]$  instead of indices  $i \in [n]$ , and the edge taken from an internal node is determined by the parity  $\bigoplus_{i \in S} x_i$  of the input (instead of the value of the single value  $x_i$  in the case of regular decision trees). The *parity decision tree (depth) complexity* of a function is the minimum depth of a parity decision tree that computes  $f$ .

Geometrically, parity decision trees correspond to partitions of the hypercube  $\{-1, 1\}^n$  into *affine subspaces*, whereas regular decision trees partition the same hypercube into subcubes. The PDT model of computation has received significant attention in recent years [MO09, ZS09, Sha11, BSK12, TWXZ13, CT14, STV14, OST<sup>+</sup>14], and in particular, there has been much interest in generalizing results that apply to normal decision trees to the more general setting of PDTs (see e.g. the survey [ZS10]).

The parity decision tree complexity of a Boolean function  $f$  can be much smaller than its regular decision tree complexity. The parity function over  $n$  variables, which can be computed by a trivial parity decision tree of depth 1 but requires regular decision tree depth  $n$ , gives the largest possible separation between the two complexity measures. As a result, many inequalities related to the decision tree complexity do not necessarily hold with respect to parity decision tree complexity. In particular, the O’Donnell–Servedio inequality does not imply that any similar inequality must hold between the sum of linear Fourier coefficients of a Boolean function and its parity decision tree complexity. Our main result shows that, nevertheless, such a generalization does hold.

**Theorem 1.** *Let  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  be computable by a parity decision tree of depth  $d$ . Define  $\sigma^2 = 4 \Pr[f(x) = 1] \Pr[f(x) = -1]$  to be the variance of  $f$ . Then*

$$\sum_{i=1}^n \widehat{f}(i) \leq \sqrt{4 \ln 2 \sigma^2 d}.$$
<sup>1</sup>

The main technical component in the proof of Theorem 1 is a fundamental inequality (presented in Lemma 3.1) concerning small-depth parity decision trees. One notable aspect about the proof of this inequality is that it is first established for a subclass of parity decision trees called *correlation-free* parity decision trees. We then show that every parity decision tree of depth  $d$  can be refined

---

<sup>1</sup>This result was originally circulated in an unpublished manuscript titled *Discrete isoperimetry via the entropy method* (2013).

to obtain a correlation-free parity decision tree of depth at most  $2d$  to obtain Lemma 3.1. See Section 3.1 for the details.

We complete the proof of Theorem 1 using an information-theoretic argument. While the proof can also be completed using analytical arguments and Jensen’s inequality, the information-theoretic argument appears to be required to obtain the sharp bounds in our theorem statement. This same argument can also be used in the regular decision tree model to sharpen the O’Donnell–Servedio theorem directly as well.

## 1.2 Application: Recursive majority function

We use Theorem 1 to obtain the first non-trivial lower bound on the parity decision tree complexity of the recursive majority function. The *3-majority function* is the function  $\text{MAJ}_3 : \{-1, 1\}^3 \rightarrow \{-1, 1\}$  defined by  $\text{MAJ}_3(x) = (-1)^{\mathbf{1}_{[x_1+x_2+x_3 < 0]}}$ . For every  $k \geq 2$ , the recursive majority function  $\text{MAJ}_3^{\otimes k} : \{-1, 1\}^{3^k} \rightarrow \{-1, 1\}$  is defined by setting

$$\text{MAJ}_3^{\otimes k}(x) = \text{MAJ}_3 \left( \text{MAJ}_3^{\otimes k-1}(x_{\{1, \dots, 3^{k-1}\}}), \text{MAJ}_3^{\otimes k-1}(x_{\{3^{k-1}+1, \dots, 2 \cdot 3^{k-1}\}}), \text{MAJ}_3^{\otimes k-1}(x_{\{2 \cdot 3^{k-1}+1, \dots, 3^k\}}) \right).$$

The recursive majority function was introduced by Boppana [SW86] to determine possible gaps between the deterministic and randomized decision tree complexity of Boolean functions. It is easy to verify that the deterministic decision tree complexity of  $\text{MAJ}_3^{\otimes k}$  is  $3^k$ . By contrast, the problem of determining the randomized decision tree complexity of  $\text{MAJ}_3^{\otimes k}$  is much more challenging: following a sequence of works on this question [SW86, JKS03, MNSX11, Leo13, MNS<sup>+</sup>13], Magniez et al. [MNS<sup>+</sup>13] have shown that the minimal depth  $R(\text{MAJ}_3^{\otimes k})$  of any randomized decision tree that computes the  $\text{MAJ}_3^{\otimes k}$  function satisfies

$$\Omega(2.57143^k) \leq R(\text{MAJ}_3^{\otimes k}) \leq O(2.64944^k)$$

but the exact randomized query complexity of the recursive majority function is still unknown.

A closely related problem that naturally arises when considering the recursive majority function is to determine its (deterministic) parity decision tree complexity. A standard adversary argument can be used to show that every parity decision tree that computes the recursive majority function has depth at least  $2^k$ . Using Theorem 1, we obtain the first lower bound on the parity decision tree complexity of the recursive majority function that improves on this trivial lower bound.

**Theorem 2.** *Every parity decision tree that computes  $\text{MAJ}_3^{\otimes k}$  has depth  $\Omega(2.25^k)$ .*

The proof of Theorem 2 is established by computing the linear Fourier coefficients of the  $\text{MAJ}_3$  function directly, using a fundamental identity on the linear Fourier coefficients of function powers (see Fact 2.7) to determine the linear Fourier coefficients of the  $\text{MAJ}_3^{\otimes k}$  function, and applying the inequality in Theorem 1. This approach is quite general, and may be useful for obtaining lower bounds on the parity decision tree complexity of other Boolean functions in the future as well.

## 2 Preliminaries

### 2.1 Information theory

All probabilities and expectations are with respect to the uniform distribution unless otherwise stated. We use boldface letters (e.g.  $\mathbf{X}$ ,  $\mathbf{x}$ ) to denote random variables. The proof of Theorem 1 uses elementary definitions and inequalities from information theory. A more thorough introduction to these tools can be found in [CT91].

**Definition 2.1.** The *entropy* of the random variable  $\mathbf{X}$  drawn from the finite sample space  $\Omega$  according to the probability mass function  $p : \Omega \rightarrow [0, 1]$  is  $H(\mathbf{X}) = -\sum_{x \in \Omega} p(x) \log p(x)$ . The *conditional entropy* of  $\mathbf{X}$  given  $\mathbf{Y}$  when they are drawn from the joint probability distribution  $p : \Omega \times \Omega' \rightarrow [0, 1]$  is  $H(\mathbf{X} \mid \mathbf{Y}) = -\sum_{x \in \Omega, y \in \Omega'} p(x, y) \log(p(y)/p(x, y))$ .

**Definition 2.2.** The *binary entropy function* is the function  $h : [0, 1] \rightarrow \mathbb{R}$  defined by  $h(t) = -t \log_2(t) - (1-t) \log_2(1-t)$ . The value  $h(t)$  represents the entropy of a random variable  $\mathbf{X}$  drawn from  $\{-1, 1\}$  with  $\Pr[\mathbf{X} = 1] = t$ .

**Fact 2.3** (Data processing inequality). *If  $\mathbf{X}$  and  $\mathbf{Z}$  are conditionally independent given  $\mathbf{Y}$ , then  $H(\mathbf{X} \mid \mathbf{Z}) \geq H(\mathbf{X} \mid \mathbf{Y})$ .*

**Fact 2.4** (Bounds on the binary entropy function). *The binary entropy function  $h : [0, 1] \rightarrow \mathbb{R}$  is bounded above and below by  $1 - t^2 \leq h(\frac{1}{2} + \frac{t}{2}) \leq 1 - \frac{t^2}{2 \ln 2}$ .*

## 2.2 Fourier analysis and function composition

We assume that the reader is familiar with the Fourier analysis of Boolean functions. For a complete introduction to the topic, see [O'D14].

**Definition 2.5.** The *composition* of  $f : \{-1, 1\}^m \rightarrow \{-1, 1\}$  and  $g : \{-1, 1\}^n \rightarrow \{-1, 1\}$  is the function  $f \circ g : \{-1, 1\}^{mn} \rightarrow \{-1, 1\}$  where

$$(f \circ g)(x) = f(g(x_1, \dots, x_n), \dots, g(x_{(m-1)n+1}, \dots, x_{mn})).$$

For  $k \geq 1$ , the  $k$ th power of  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  is the function  $f^{\otimes k} : \{-1, 1\}^{n^k} \rightarrow \{-1, 1\}$  defined recursively by setting  $f^{\otimes 1} = f$  and  $f^{\otimes k} = f \circ f^{\otimes k-1}$ .

**Remark 2.6.** As we can verify directly, the recursive majority function  $\text{MAJ}_3^{\otimes k}$  is the  $k$ th power of the  $\text{MAJ}_3$  function.

We use the following fact on the linear Fourier coefficients of composed functions. (See Appendix A for a proof of this fact.)

**Fact 2.7.** *For any  $f : \{-1, 1\}^m \rightarrow \{-1, 1\}$  and any balanced function  $g : \{-1, 1\}^n \rightarrow \{-1, 1\}$ ,*

$$\sum_{k \in [mn]} \widehat{f \circ g}(k) = \left( \sum_{i \in [n]} \widehat{f}(i) \right) \left( \sum_{j \in [m]} \widehat{g}(j) \right).$$

## 2.3 Parity decision trees

As mentioned in the introduction, a parity decision tree is a rooted full binary tree where each internal node is associated with a set  $S \subseteq [n]$ , the two edges leading to the children of a node are labelled with  $-1$  and  $1$ , respectively, and each leaf is associated with a value in  $\{-1, 1\}$ . Each input  $x \in \{-1, 1\}^n$  defines a path to a unique leaf in a parity decision tree  $T$  by following the edge labelled with  $\chi_S(x)$  from a node labelled with  $S$ . We say that the tree  $T$  *computes* the Boolean function  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  if each input  $x$  defines a path in  $T$  to a leaf labelled with  $f(x)$ . When  $T$  computes  $f$  and  $\ell$  is a leaf of  $T$ , we write  $f(\ell)$  to denote the label of  $\ell$ .

We can represent each leaf of a parity decision tree  $T$  with a vector  $\ell \in \{-1, 0, 1\}^n$  where  $\ell_i$  is the expected value of the coordinate  $x_i$  over the uniform distribution of all inputs  $x \in \{-1, 1\}^n$  that define a path to the leaf  $\ell$  in  $T$ . We let  $\text{leaf}_T : \{-1, 1\}^n \rightarrow \{-1, 0, 1\}^n$  be the function that returns the vector representation of the leaf reached by the path defined in  $T$  for every input  $x \in \{-1, 1\}^n$ .

### 3 Proof of Theorem 1

The main technical component of the proof of Theorem 1 is the following inequality.

**Lemma 3.1.** *For any parity decision tree  $T$  of depth  $d$ ,  $\mathbf{E}_{\ell \in T} [(\sum_{i=1}^n \ell_i)^2] \leq 2d$ .*

We now complete the proof of Theorem 1 assuming Lemma 3.1. The proof of the lemma then follows in the next subsection.

**Theorem 1 (Restated).** *Let  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  be computable by a parity decision tree  $T$  of depth  $d$ . Define  $\sigma^2 = 4 \Pr[f(x) = 1] \Pr[f(x) = -1]$  to be the variance of  $f$ . Then*

$$\sum_{i=1}^n \widehat{f}(i) \leq \sqrt{4 \ln 2 \sigma^2 d}.$$

*Proof.* Draw  $\mathbf{X} \in \{-1, 1\}^n$  and  $\mathbf{i} \in [n]$  independently and uniformly at random. Let us first compute the conditional entropy  $H(\mathbf{X}_{\mathbf{i}} \mid f(\mathbf{X}))$ . Write  $\mu = \Pr[f(\mathbf{X}) = 1]$ . Then

$$\Pr_{\mathbf{X}}[\mathbf{X}_{\mathbf{i}} = 1 \mid f(\mathbf{X}) = 1] = \frac{\mathbf{E}[(\frac{1+\mathbf{X}_{\mathbf{i}}}{2})(\frac{1+f(\mathbf{X})}{2})]}{\Pr[f(\mathbf{X}) = 1]} = \frac{1}{2} + \frac{\widehat{f}(i)}{4\mu}$$

and so

$$\Pr_{\mathbf{X}, \mathbf{i}}[\mathbf{X}_{\mathbf{i}} = 1 \mid f(\mathbf{X}) = 1] = \frac{1}{2} + \sum_{i=1}^n \frac{\widehat{f}(i)}{4\mu n}.$$

Similarly,

$$\Pr_{\mathbf{X}, \mathbf{i}}[\mathbf{X}_{\mathbf{i}} = 1 \mid f(\mathbf{X}) = -1] = \frac{1}{2} - \sum_{i=1}^n \frac{\widehat{f}(i)}{4(1-\mu)n}.$$

By the definition of conditional entropy and the upper bound in Fact 2.4,

$$\begin{aligned} H(\mathbf{X}_{\mathbf{i}} \mid f(\mathbf{X})) &= \mu h\left(\frac{1}{2} + \sum_{i=1}^n \frac{\widehat{f}(i)}{4\mu n}\right) + (1-\mu) h\left(\frac{1}{2} + \sum_{i=1}^n \frac{\widehat{f}(i)}{4(1-\mu)n}\right) \\ &\leq \mu \left(1 - \frac{1}{2 \ln 2} \left(\frac{\sum_i \widehat{f}(i)}{2\mu n}\right)^2\right) + (1-\mu) \left(1 - \frac{1}{2 \ln 2} \left(\frac{\sum_i \widehat{f}(i)}{2(1-\mu)n}\right)^2\right) \\ &= 1 - \frac{(\sum_i \widehat{f}(i))^2}{8 \ln 2 \mu(1-\mu)n^2}. \end{aligned} \tag{1}$$

Since the leaf reached in  $T$  by an input  $x$  determines  $f(x)$ , the data processing inequality implies that:

$$H(\mathbf{X}_{\mathbf{i}} \mid f(\mathbf{X})) \geq H(\mathbf{X}_{\mathbf{i}} \mid \text{leaf}_T(\mathbf{X})). \tag{2}$$

We also have that

$$H(\mathbf{X}_{\mathbf{i}} \mid \text{leaf}_T(\mathbf{X})) = \mathbf{E}_{\ell \in T} \left[ h\left(\frac{1}{2} + \frac{\sum_i \ell_i}{2n}\right) \right]$$

where the expectation is over the distribution defined by the relative mass of each leaf in  $T$ . Applying the lower bound in Fact 2.4, we get

$$H(\mathbf{X}_{\mathbf{i}} \mid \text{leaf}_T(\mathbf{X})) \geq 1 - \mathbf{E}_{\ell \in T} \left[ \left(\frac{\sum_i \ell_i}{2n}\right)^2 \right]. \tag{3}$$

Combining (1)–(3), we obtain

$$\left(\sum_i \widehat{f}(i)\right)^2 \leq 2 \ln 2 \cdot 4\mu(1-\mu) \mathbf{E}_\ell \left[ \left(\sum_i \ell_i\right)^2 \right]$$

and the theorem follows from the bound in Lemma 3.1.  $\square$

**Remark 3.2.** A result that is similar to Theorem 1, but with a slightly weaker bound, can also be obtained directly from Lemma 3.1 and Jensen’s inequality. This approach gives the weaker bound  $\sum_{i=1}^n \widehat{f}(i) \leq \sqrt{2d}$ . See Appendix B for the details.

### 3.1 Proof of Lemma 3.1

The proof of Lemma 3.1 has three main components. The first is a proof of the lemma for a class of parity decision trees that we call *(pairwise) correlation-free*.

**Definition 3.3.** The parity decision tree  $T$  is *(pairwise) correlation-free* when for every  $i \neq j \in [n]$  and any path in the tree  $T$ , if  $x_i \oplus x_j$  is fixed by the queries in the path, then so are  $x_i$  and  $x_j$ .

**Proposition 3.4.** Let  $T$  be a correlation-free parity decision tree of depth  $d$ . Then  $\mathbf{E}(\sum_i \ell_i)^2 \leq d$ .

*Proof.* Consider any node  $v$  in the parity decision tree that fixes the parity  $x_i \oplus x_j$ . Since  $T$  is correlation-free, every leaf below  $v$  satisfies  $\ell_i, \ell_j \neq 0$ . In particular,  $\Pr_{\ell \sim v}[\ell_i \ell_j = -1] = \Pr_{\ell \sim v}[\ell_i \ell_j = 1] = 1/2$  so  $\mathbf{E}_{\ell \sim v} \ell_i \ell_j = 0$ . And every path that reaches a leaf without fixing  $x_i \oplus x_j$  does not set both  $x_i$  and  $x_j$ , so such a leaf  $\ell$  satisfies  $\ell_i \ell_j = 0$ . This means that for every  $i \neq j$ ,  $\mathbf{E} \ell_i \ell_j = 0$  and so

$$\mathbf{E}(\sum_i \ell_i)^2 = \sum_i \mathbf{E}(\ell_i)^2 + \sum_{i \neq j} \mathbf{E} \ell_i \ell_j \leq d, \quad (4)$$

where the final inequality uses the fact that at most  $d$  coordinates can be fixed by the queries of any path in  $T$ .  $\square$

We want to use Proposition 3.4 by showing that we can refine every parity decision tree into an uncorrelated parity decision tree without increasing its depth by too much. The following proposition formalizes this statement.

**Proposition 3.5.** Let  $T$  be a parity decision tree of depth  $d$ . Then there is a refinement  $T'$  of  $T$  which is an uncorrelated parity decision tree of depth at most  $2d$ .

*Proof.* For each leaf of  $T$ , let  $J$  be a set of disjoint pairs  $(i, j)$  of coordinates such that  $x_i \oplus x_j$  is fixed but neither  $x_i$  nor  $x_j$  have been fixed by the queries down the path to the leaf. Refine  $T$  by querying the first coordinate in each such pair. Once we have done this refinement at every leaf, the resulting tree is uncorrelated. To complete the proof of the proposition, it remains to show that at most  $2d$  disjoint pairs of correlated coordinates can occur in any path on the tree  $T$ .

Let  $V$  be the subspace of  $\{0, 1\}^n$  spanned by the (at most)  $d$  queries down any fixed path in  $T$ . Let  $S$  be a maximal linearly independent subset of  $V$  containing only vectors of Hamming weight 1 or 2. Since  $V$  is a  $d$ -dimensional subspace,  $|S| \leq d$ . Let  $J$  be the set of coordinates that are set to 1 in at least one vector in  $S$ . Then  $|J| \leq 2d$ . Furthermore, if  $i$  is fixed or correlated, there exists a vector  $v$  of Hamming weight at most 2 in  $V$  for which  $v_i = 1$ . This means that either  $v \in S$  or  $v$  is a linear combination of some vectors in  $S$ ; either case implies that  $i \in J$ .  $\square$

The third and final component of our proof of the lemma is a simple argument showing that refining a decision tree can only increase the value of  $\mathbf{E}(\sum_i \ell_i)^2$ .

**Proposition 3.6.** *Let  $T'$  be any refinement of the parity decision tree  $T$ . Then*

$$\mathbf{E}_{\ell \in T} \left( \sum_{i=1}^n \ell_i \right)^2 \leq \mathbf{E}_{\ell' \in T'} \left( \sum_{i=1}^n \ell'_i \right)^2.$$

*Proof.* It suffices to establish the proof in the case where  $T'$  replaces one leaf of  $T$  with an extra node. Let  $v$  be the leaf in  $T$  that we replace with the node with leaves  $u, w$ . Let  $\rho$  be the probability that a random input  $x$  reaches the leaf  $v$  in  $T$ . Then

$$\mathbf{E}_{\ell' \in T'} \left( \sum_{i=1}^n \ell'_i \right)^2 - \mathbf{E}_{\ell \in T} \left( \sum_{i=1}^n \ell_i \right)^2 = \rho \cdot \left( \frac{1}{2} \left( \sum_{i=1}^n u_i \right)^2 + \frac{1}{2} \left( \sum_{i=1}^n w_i \right)^2 - \left( \sum_{i=1}^n v_i \right)^2 \right).$$

Let  $S \subseteq [n]$  be the set of coordinates that are fixed by the query at the node that replaced  $v$ . Then  $v_i = 0$  for each  $i \in S$ , and  $\delta := \sum_{i \in S} u_i = -\sum_{i \in S} w_i$ . Write  $\gamma = \sum_i v_i$ . Then

$$\frac{1}{2} \left( \sum_{i=1}^n u_i \right)^2 + \frac{1}{2} \left( \sum_{i=1}^n w_i \right)^2 - \left( \sum_{i=1}^n v_i \right)^2 = \frac{1}{2}(\gamma + \delta)^2 + \frac{1}{2}(\gamma - \delta)^2 - \gamma^2 = \delta^2 \geq 0. \quad \square$$

We can now complete the proof of the lemma.

*Proof of Lemma 3.1.* Let  $T'$  be the uncorrelated parity decision tree of depth at most  $2d$  obtained by refining  $T$ , as promised by Proposition 3.5. By Propositions 3.6 and 3.4,

$$\mathbf{E}_{\ell \in T} \left( \sum_{i=1}^n \ell_i \right)^2 \leq \mathbf{E}_{\ell' \in T'} \left( \sum_{i=1}^n \ell'_i \right)^2 \leq 2d. \quad \square$$

**Remark 3.7.** The same arguments in the proof of Lemma 3.1 can also be sharpened to show that the expression  $\mathbf{E}(\sum_i \ell_i)^2$  is bounded above by 2 times the average depth of the parity decision tree  $T$ .

**Remark 3.8.** When  $T$  is a standard decision tree, (4) directly implies that  $\mathbf{E}(\sum_i \ell_i)^2 \leq d$ . It is natural to ask whether Lemma 3.1 can be sharpened to obtain the same bound for parity decision trees as well. It cannot: consider the  $\text{MAJ}_3 : \{-1, 1\}^3 \rightarrow \{-1, 1\}$  function, which returns the sign of  $x_1 + x_2 + x_3$ . One parity decision tree that computes  $\text{MAJ}_3$  queries  $x_1 x_2$  at the root and then queries  $x_1$  if  $x_1 x_2 = 1$ , or  $x_3$  otherwise. This tree has depth 2 but  $\mathbf{E}(\sum_i \ell_i)^2 = \frac{5}{2} > 2$ .

## 4 The recursive majority function

Let us now see how Theorem 1 yields a lower bound on the parity decision tree complexity of the recursive majority function.

**Theorem 2** (Restated). *Every parity decision tree that computes  $\text{MAJ}_3^{\otimes k}$  has depth  $\Omega(2.25^k)$ .*

*Proof.* By direct calculation, we observe that the Fourier expansion of the  $\text{MAJ}_3$  function is

$$\text{MAJ}_3(x_1, x_2, x_3) = \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3 - \frac{1}{2}x_1 x_2 x_3.$$

By Fact 2.7, for every  $k > 1$  we have

$$\sum_{i \in [3^k]} \widehat{\text{MAJ}_3^{\otimes k}}(i) = \left( \sum_{i \in [3]} \widehat{\text{MAJ}_3}(i) \right) \left( \sum_{j \in [3^{k-1}]} \widehat{\text{MAJ}_3^{\otimes k-1}}(j) \right) = \frac{3}{2} \left( \sum_{j \in [3^{k-1}]} \widehat{\text{MAJ}_3^{\otimes k-1}}(j) \right).$$

By induction, this identity yields

$$\sum_{i \in [3^k]} \widehat{\text{MAJ}_3^{\otimes k}}(i) = \left( \frac{3}{2} \right)^k.$$

Let  $d$  be the minimal depth of any parity decision tree that computes  $\text{MAJ}_3^{\otimes k}$ . By Theorem 1, we have  $(\frac{3}{2})^k \leq \sqrt{4 \ln 2 d}$  and so  $d \geq \Omega((\frac{3}{2})^{2k}) = \Omega(2.25^k)$ .  $\square$

## 5 Conclusion and open problem

We have shown that the O’Donnell–Servedio inequality generalizes to the setting of parity decision trees. A related conjecture of Parikshit Gopalan and Rocco Servedio posits that the O’Donnell–Servedio inequality can also be generalized in a different direction as well, to the setting of Boolean functions with low *Fourier degree*, where the Fourier degree of a Boolean function  $f$  is the size of the largest set  $S$  such that  $\widehat{f}(S) \neq 0$ .

**Gopalan–Servedio Conjecture [O’D12].** *Let  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  be a Boolean function with Fourier degree  $d$ . Then  $\sum_{i=1}^n \widehat{f}(i) \leq O(\sqrt{d})$ .*

While the Gopalan–Servedio conjecture and Theorem 1 both generalize the O’Donnell–Servedio inequality (as Fourier degree and parity decision tree depth are both upper bounded by regular decision tree depth), they are incomparable to each other — the  $n$ -variable parity function has PDT depth 1 and Fourier degree  $n$ , and conversely there are functions whose PDT depth is polynomially larger than its Fourier degree [OST<sup>+</sup>14].

## Acknowledgements

We thank Ryan O’Donnell and Rocco Servedio for insightful conversations. We also thank the anonymous referees of an earlier version of this manuscript for valuable feedback.

## References

- [Blu03] Avrim Blum. Machine learning: a tour through some favorite results, directions, and open problems. FOCS 2003 tutorial slides, available at <http://www-2.cs.cmu.edu/~avrim/Talks/FOCS03/tutorial.ppt>, 2003.
- [BSK12] Eli Ben-Sasson and Swastik Kopparty. Affine dispersers from subspace polynomials. *SIAM Journal on Computing*, 41(4):880–914, 2012.
- [CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 1991.
- [CT14] Gil Cohen and Avishay Tal. Two structural results for low degree polynomials and applications. *arXiv preprint*, 1404.0654, 2014.

- [JKS03] T. S. Jayram, Ravi Kumar, and D. Sivakumar. Two applications of information complexity. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA*, pages 673–682, 2003.
- [Leo13] Nikos Leonardos. An improved lower bound for the randomized decision tree complexity of recursive majority,. In *Proceedings of Automata, Languages, and Programming - 40th International Colloquium, ICALP 2013, Part I*, pages 696–708, 2013.
- [MNS<sup>+</sup>13] Frédéric Magniez, Ashwin Nayak, Miklos Santha, Jonah Sherman, Gábor Tardos, and David Xiao. Improved bounds for the randomized decision tree complexity of recursive majority. *arXiv preprint*, 1309.7565, 2013.
- [MNSX11] Frédéric Magniez, Ashwin Nayak, Miklos Santha, and David Xiao. Improved bounds for the randomized decision tree complexity of recursive majority. In *Automata, Languages and Programming - 38th International Colloquium, ICALP 2011, Zurich, Switzerland, July 4-8, 2011, Proceedings, Part I*, pages 317–329, 2011.
- [MO09] Ashley Montanaro and Tobias Osborne. On the communication complexity of xor functions. *arXiv preprint*, 0909.3392, 2009.
- [O’D12] Ryan O’Donnell. Open problems in analysis of Boolean functions. *arXiv preprint*, 1204.6447, 2012.
- [O’D14] Ryan O’Donnell. *Analysis of Boolean Functions*. Cambridge University Press (available online at <http://analysisofbooleanfunctions.org>), 2014.
- [OS08] Ryan O’Donnell and Rocco Servedio. Learning monotone decision trees in polynomial time. *SIAM Journal on Computing*, 37(3):827–844, 2008.
- [OST<sup>+</sup>14] Ryan O’Donnell, Xiaorui Sun, Li-Yang Tan, John Wright, and Yu Zhao. A composition theorem for parity kill number. In *IEEE 29th Conference on Computational Complexity, CCC 2014, Vancouver*, pages 144–154, 2014.
- [Sha11] Ronen Shaltiel. Dispersers for affine sources with sub-polynomial entropy. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 247–256. IEEE, 2011.
- [STV14] Amir Shpilka, Avishay Tal, and Ben Lee Volk. On the structure of boolean functions with small spectral norm. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 37–48. ACM, 2014.
- [SW86] Michael E. Saks and Avi Wigderson. Probabilistic boolean decision trees and the complexity of evaluating game trees. In *27th Annual Symposium on Foundations of Computer Science, Toronto, Canada, 27-29 October 1986*, pages 29–38, 1986.
- [TWXZ13] Hing Yin Tsang, Chung Hoi Wong, Ning Xie, and Shengyu Zhang. Fourier sparsity, spectral norm, and the log-rank conjecture. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 658–667. IEEE, 2013.
- [ZS09] Zhiqiang Zhang and Yaoyun Shi. Communication complexities of symmetric xor functions. *Quantum Information & Computation*, 9(3):255–263, 2009.
- [ZS10] Zhiqiang Zhang and Yaoyun Shi. On the parity complexity measures of boolean functions. *Theoretical Computer Science*, 411(26):2612–2618, 2010.

## A Multiplicativity of the level-1 Fourier mass

Fact 2.7 is a direct consequence of the following identity.

**Proposition A.1.** *For any function  $f : \{-1, 1\}^m \rightarrow \{-1, 1\}$ , any balanced function  $g : \{-1, 1\}^n \rightarrow \{-1, 1\}$ , and any  $i \in [m]$  and  $j \in [n]$ ,*

$$\widehat{f \circ g}((i-1)n+j) = \widehat{f}(i)\widehat{g}(j).$$

*Proof.* By definition, the Fourier expansion of  $f$ , and linearity of expectation,

$$\begin{aligned} \widehat{f \circ g}((i-1)n+j) &= \mathbf{E}_x [f(g(x_1, \dots, x_n), \dots, g(x_{(m-1)n+1}, \dots, x_{mn})) \cdot x_{(i-1)n+j}] \\ &= \sum_{S \subseteq [n]} \widehat{f}(S) \mathbf{E}_x \left[ \prod_{k \in S} g(x_{(k-1)n+1}, \dots, x_{kn}) \cdot x_{(i-1)n+j} \right]. \end{aligned} \quad (5)$$

When  $i \notin S$ ,

$$\mathbf{E}_x \left[ \prod_{k \in S} g(x_{(k-1)n+1}, \dots, x_{kn}) \cdot x_{(i-1)n+j} \right] = \mathbf{E}_x \left[ \prod_{k \in S} g(x_{(k-1)n+1}, \dots, x_{kn}) \right] \cdot \mathbf{E}_x [x_{(i-1)n+j}] = 0.$$

Similarly, when  $S \setminus \{i\} \neq \emptyset$ , we can fix any  $\ell \in S \setminus \{i\}$  and observe that

$$\begin{aligned} \mathbf{E}_x \left[ \prod_{k \in S} g(x_{(k-1)n+1}, \dots, x_{kn}) \cdot x_{(i-1)n+j} \right] \\ = \mathbf{E}_x [g(x_{(\ell-1)n+1}, \dots, x_{\ell n})] \cdot \mathbf{E}_x \left[ \prod_{k \in S \setminus \{\ell\}} g(x_{(k-1)n+1}, \dots, x_{kn}) \cdot x_{(i-1)n+j} \right]. \end{aligned}$$

When  $g$  is balanced,  $\mathbf{E}_x [g(x_{(\ell-1)n+1}, \dots, x_{\ell n})] = 0$  so the only non-zero term of the sum in (5) is the one where  $S = \{i\}$  and

$$\begin{aligned} \widehat{f \circ g}((i-1)n+j) &= \widehat{f}(i) \mathbf{E}_x [g(x_{(i-1)n+1}, \dots, x_{in}) x_{(i-1)n+j}] \\ &= \widehat{f}(i) \mathbf{E}_x [g(x_1, \dots, x_n) x_j] \\ &= \widehat{f}(i)\widehat{g}(j). \end{aligned} \quad \square$$

## B Coarser bounds

We can obtain a weaker version of Theorem 1 by combining Lemma 3.1 with the following easy inequality which is essentially equivalent to Lemma 3 in [OS08].

**Lemma B.1** (O'Donnell and Servedio [OS08]). *Let  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  be computable by a parity decision  $T$ . Then*

$$\sum_{i=1}^n \widehat{f}(i) < \mathbf{E}_{\ell \in T} \left[ \left| \sum_{i=1}^n \ell_i \right| \right].$$

*Proof.* The linear Fourier coefficients of  $f$  satisfy

$$\widehat{f}(i) = \mathbf{E}_x [f(x) x_i] = \mathbf{E}_{\ell \in T} \mathbf{E}_{x:t(x)=\ell} [f(x) x_i] = \mathbf{E}_{\ell \in T} [f(\ell) \mathbf{E}_{x:t(x)=\ell} [x_i]] = \mathbf{E}_{\ell \in T} [f(\ell) \ell_i].$$

So  $\sum_i \widehat{f}(i) = \mathbf{E}_{\ell \in T} [f(\ell) \sum_i \ell_i] \leq \mathbf{E}_{\ell \in T} [|\sum_i \ell_i|]$ .  $\square$

We are now ready to complete the proof of the slightly weaker version of Theorem 1.

**Theorem 3.** *Let  $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$  be computable by a parity decision tree of depth  $d$ . Define  $\sigma^2 = 4 \Pr[f(x) = 1] \Pr[f(x) = -1]$  to be the variance of  $f$ . Then*

$$\sum_{i=1}^n \widehat{f}(i) \leq \sqrt{2d}.$$

*Proof.* By Lemma B.1 and Jensen's inequality,

$$\left( \sum_{i=1}^n \widehat{f}(i) \right)^2 \leq \mathbf{E}_{\ell \in T} \left[ \left| \sum_{i=1}^n \ell_i \right| \right]^2 \leq \mathbf{E}_{\ell \in T} \left[ \left( \sum_{i=1}^n \ell_i \right)^2 \right].$$

Theorem 1 then follows directly from Lemma 3.1. □